Forecasting Urban Air Quality in Makassar Using ARIMA and ARIMAX Models.

Dimas Pramono[1], *M.* Rahmatulloh[1], *Muh.* Baidlowi[1], and *Pebri* Putra[1]

[1]Informatics Engineering, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia

Abstract. This study presents a comparative analysis of the ARIMA and ARIMAX models to forecast the daily PM2.5 air pollutant concentration in Makassar, Indonesia. Air quality data were collected from AQICN, while meteorological parameters (temperature, humidity, and wind speed) were obtained from BMKG. Initial preprocessing included cleaning missing values using hybrid mean interpolation and converting the data into time series format. The ARIMA (0,1,2) model was selected as the optimal univariate approach based on performance metrics such as RMSE, MAPE, and Ljung-Box test. An ARIMAX (1,1,0) model was constructed by integrating exogenous meteorological variables, demonstrating slightly improved performance. The results confirm that meteorological factors contribute to PM2.5 fluctuations and that ARIMAX offers a more representative model for urban air quality forecasting.

Keywords: Air quality, PM2.5, ARIMA, ARIMAX, Forecasting, Time series, Meteorology

[1] Corresponding author: mf305456@email.org

1 Introduction

Air quality plays a critical role in assessing environmental health, especially in growing urban areas. Makassar, a major city in Indonesia, is experiencing increased industrial activity, vehicular traffic, and population growth, resulting in higher emissions of air pollutants. One of the most concerning pollutants is PM2.5, airborne particles smaller than 2.5 micrometers that pose significant health risks due to their ability to penetrate deep into the respiratory system.

Given the rising concerns over pollution levels, particularly the air quality index (AQI) reaching unhealthy thresholds in areas such as Daya and Malengkeri terminals, predictive modeling is essential. Forecasting tools provide insights into future trends, enabling policymakers to implement mitigation strategies. This study utilizes ARIMA and ARIMAX models to forecast PM2.5 levels, with the latter incorporating meteorological factors for improved accuracy.

2 Literature Review

Muzakki et al. (2024) compared ARIMA and ARIMAX models for AQI forecasting in Central Jakarta and concluded that ARIMAX (1,1,1) outperforms ARIMA (1,1,1) when meteorological variables are included. Additional references from WHO (2021), Jian et al. (2012), and Cobourn (2010) affirm that average temperature, humidity, and wind speed significantly affect pollutant dispersion. These studies lay the groundwork for using ARIMAX in Makassar.

3 Methodology

3.1 Data Collection and Preprocessing

Two datasets were used in this study: (1) daily PM2.5 concentrations retrieved from the AQICN platform using an API from monitoring stations in Makassar, and (2) meteorological data including daily mean temperature (TAVG), relative humidity (RH_AVG), and wind speed (FF_AVG) collected from BMKG. The observation period spans from February 7 to May 24, 2025.

Missing values were handled using a hybrid mean interpolation method. This combines simple averaging and linear interpolation based on the position and frequency of missing entries, ensuring continuity in time series data.

3.1.1 Research Method

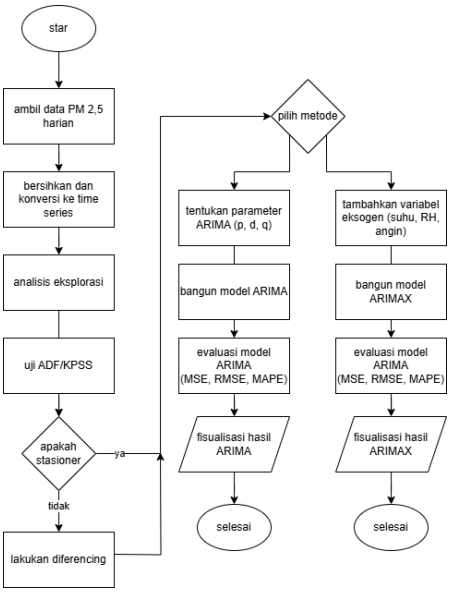

Fig. 1. Flowchart Research

3.1.2 Datasets PM2.5 Makassar

Tabel 1. Datasets PM2.5 Makassar

TANGGAL	PM2.5						
	Min	Max	Median	Q1	Q3	Standar Deviase	Count
Friday, 7 February 2025	5.5	6.81	6.04	5.72	6.51	0.417	21
Saturday, 8 February 2025	5.25	7.83	6.13	5.69	7.07	0.765	24
Sunday, 9 February 2025	7.58	8.21	7.89	7.81	7.97	0.146	24
Monday, 10 February 2025	7.58	8.75	8.07	7.82	8.5	0.38	24
Tuesday, 11 February 2025	7.93	10.91	9.44	8.76	10.58	1.006	23
Wednesday, 12 February 2025	9.78	11.56	10.19	10.07	10.76	0.528	24
Thursday, 13 February 2025	8.19	10.81	9.82	8.91	10.57	0.893	24
Friday, 14 February 2025	8	9.94	8.52	8.44	9.22	0.539	24
Saturday, 15 February 2025	8.56	10.06	9.03	8.77	9.31	0.435	24
Sunday, 16 February 2025	8.73	12.44	9.61	9.38	10.43	1.076	24
Monday, 17 February 2025	11.46	12.94	12.65	11.95	12.8	0.468	24

3.1.3 Datasets Climate and Meteorology of Makassar

Tabel 2. Datasets Climate and Meteorology of Makassar

TANGGAL	TN	TX	TAVG	RH_AVG	RR	SS	FF_X	DDD_X	FF_AVG	DDD_CAR
01/02/2025	24	31	26.6	87	27.8	2	6	30	2	E
02/02/2025	25	31	27.8	82	9.6	5	5	320	2	N
03/02/2025	26	32	28.7	73	0	8	3	290	1	NW
04/02/2025	26	31	28	80	0	5	5	330	1	NW
05/02/2025	26	31	28	80	0.6	3	5	300	2	NW
06/02/2025	25	30	26.5	87	10.4	5	6	300	2	SE
07/02/2025	24	27	25.1	92	33.6	0	7	310	3	NE
08/02/2025	24	29	26	91	77.2	0	8	280	5	W
09/02/2025	24	29	24	96	44	4	8	280	3	E
10/02/2025	23	30	27.2	89	89	0	9	290	6	W
11/02/2025	25	29	25.7	94	28	2	8	280	3	E
12/02/2025	24	30	26	91	99.2	0	8	310	2	E
13/02/2025	24	31	27.2	87	7	1	5	310	2	E

3.1.4 Handling Missing Data

In the process of preparing the dataset, missing values were identified in both PM2.5 and meteorological variables. Missing values are a critical issue in time series analysis, as they can lead to biased parameter estimation and compromise model accuracy if not handled properly.
To address this issue, a method called Hybrid Mean Interpolation was employed.

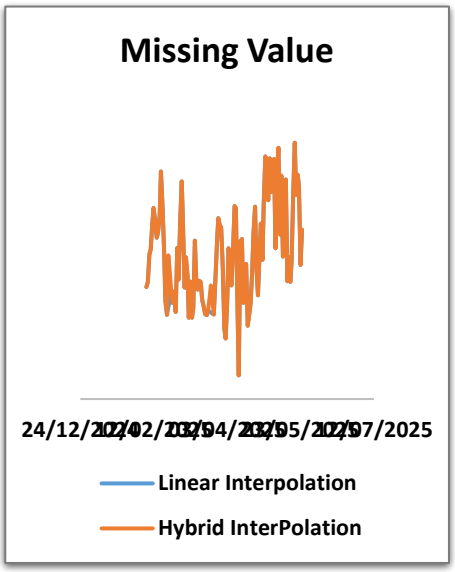

Fig. 2. Missing Data Visualization

This method combines simple linear interpolation with localized mean estimation to predict missing values based on surrounding observations. It was chosen for its balance between computational efficiency and its ability to preserve the natural variation in time series data.

Reasons for Choosing Hybrid Mean Interpolation: Compared to pure linear interpolation, it avoids overly sharp estimations in fluctuating data. It is computationally efficient and easy to implement. Unlike regression or forecasting-based imputation, it does not require additional model training or complexity.

However, one limitation is that it may produce unreliable estimates when large blocks of data are missing. Thus, this method was applied only to non-sequential and limited missing entries to maintain data integrity.

3.2 ARIMA Modeling

ARIMA (AutoRegressive Integrated Moving Average) models are standard tools for univariate time series forecasting. The ARIMA(p,d,q) structure requires the data to be stationary. Augmented Dickey-Fuller (ADF) tests were conducted, followed by differencing when necessary. ACF and PACF plots were used to determine the appropriate values of p and q. Model performance was evaluated using RMSE, MAPE, and Bayesian Information Criterion (BIC), with Ljung-Box Q-tests confirming residual randomness.

Tabel 3. Model Arima

Model ARIMA	Error			Variabel Signifikan	Variabel Non-Signifikan	Ujung Box Q-Test	
	MAPE	RMSE	N-BIC			P Value	White-Noise
(0,1,1)	27.709	2.243	1.703	1	0	0.049	White-Noise

(0,1,2)	25.004	2.078	1.595	1	0	0.542	No
(1,1,0)	27.397	2.308	1.761	0	1	0.002	White-Noise
(1,1,1)	26.76	2.124	1.638	2	1	0.114	No
(1,1,2)	25.028	2.087	1.647	1	2	0.545	No
(2,1,0)	26.165	2.167	1.679	0	2	0.297	No
(2,1,1)	26.418	2.17	1.725	1	2	0.285	No
(2,1,2)	26.427	2.118	1.721	2	2	0.117	No

There is equation of Model ARIMA:

$$(1 - B)y_t = (1 + \theta_1 B + \theta_1 B2)\varepsilon_t \quad (1)$$

3.3 ARIMAX Modeling

To improve prediction accuracy, ARIMAX extends ARIMA by incorporating exogenous predictors. Meteorological parameters were aligned temporally with PM2.5 data and normalized where needed. The ARIMAX model was trained using 80% of the dataset, with 20% reserved for validation. Regression coefficients were estimated, and model accuracy was evaluated using MSE, RMSE, and MAPE.

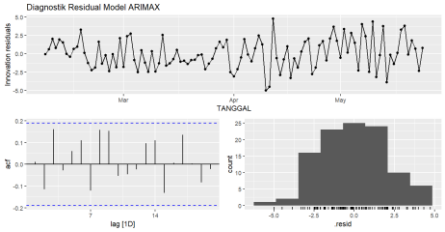

Fig. 3. Evaluate Datasets with Residual

3.4 Tools

Data processing and modeling were conducted using R Studio, SPSS, and Microsoft Excel.

4 Results and Discussion

4.1 ARIMA Model Performance

Among several tested configurations, the ARIMA(0,1,2) model achieved the best performance. Model equation:

$$y_t - y_{t\_1} = \varepsilon_t + 0.290\varepsilon_{t\_1} + 0.501\varepsilon_{t\_2} \quad (2)$$

Where yt is the PM2.5 value, and εt is the error term. Evaluation metrics:

Tabel 4. Evaluation Model ARIMA

Model	MSE	RMSE	MAPE (%)	MAE	N-BIC	Ljung-Box Q-Test
ARIMA (0,1,2)	4.318	2.078	25.004	1.656	1.595	0.542

This indicates a reasonably well-fitting model with predictive capacity for short-term forecasts.

4.2 ARIMAX Model Performance

The ARIMAX(1,1,0) model incorporated TAVG, RH_AVG, and FF_AVG as exogenous variables. Performance metrics:

While the improvement is marginal, the inclusion of meteorological variables adds contextual understanding and slightly improves forecast reliability. Regression analysis confirmed significant contributions from weather factors to PM2.5 variations.

Tabel 5. Evaluation Model ARIMAX

Model	MSE	RMSE	MAPE (%)	MAE	N-BIC	Ljung-Box Q-Test
ARIMAX (1,1,0)	4.318	3.46	23.8	2.76	1.595	0.542

This indicates a reasonably well-fitting model with predictive capacity for short-term forecasts.

4.3 Visualization and Comparative Analysis

Forecasts from both models were plotted against actual PM2.5 values. ARIMA predictions follow historical trends accurately but do not account for sudden meteorological changes. ARIMAX predictions track these shifts more closely, especially during weather-induced anomalies. In summary:

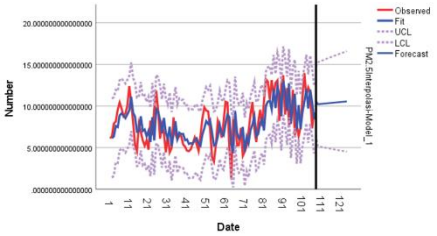

Fig. 4. ARIMA Visualization

Slightly better, meteorology-aware of ARIMAX visualization in the bottom here:

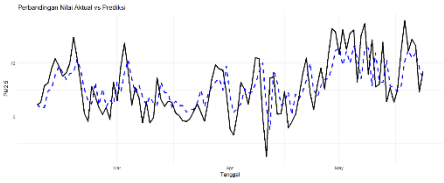

Fig. 5. ARIMAX Visualization

5 Conclusion

Following Muzakki et al. (2024), the ARIMA model captures underlying PM2.5 trends, but fails to reflect daily fluctuations caused by weather. ARIMAX, incorporating TAVG, RH_AVG, and FF_AVG, adapts more dynamically.

Tabel 6. Evaluation Comparison

Model	RMSE	MAPE (%)	MAE	AIC	BIC	Notes
ARIMA (0,1,2)	2.078	25.004	1.655	6366.21	6379.98	Univariate, Stable
ARIMAX (1,1,0)	4.318	23.8	2.76	6311.82	6339.37	Meteorological Factors Included

This study concludes that:

The ARIMA(0,1,2) model provides reliable baseline forecasts for PM2.5 levels.

The ARIMAX(1,1,0) model shows improved performance by incorporating temperature, humidity, and wind speed.

Meteorological factors significantly influence PM2.5 concentrations in Makassar.

Visualization and evaluation metrics confirm the suitability of both models for air quality forecasting, with ARIMAX being more representative in multi-variable scenarios.

6 References

[1] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, (1976).

[2] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, 2nd ed., OTexts (2018). https://otexts.com/fpp2/.

[3] C.A. Pope III, D.W. Dockery, Health effects of fine particulate air pollution: Lines that connect. J. Air Waste Manag. Assoc., 56, 709–742 (2006). https://doi.org/10.1080/10473289.2006.10464485

[4] World Health Organization, WHO Global Air Quality Guidelines: PM2.5, PM10, NO2, SO2, CO, O3 (2021).

[5] S. Wahid, H. Setyawan, Peramalan Data Deret Waktu Menggunakan Model ARIMA. Indonesian J. Applied Statistics, 3(2), 136–147 (2020).

[6] Jurnal JTIK (2024), Forecasting the air quality index using ARIMAX. J. Teknol. Informasi dan Komunikasi, 8(1).

[7] A. Ramdhan, S. Syamsuddin, Urban Growth and Industrial Development in Makassar City. J. Environ. Manage. Tourism, 11(4), 789–796 (2020).

[8] M. Akbar, Repository ITK, Stationarity in Inflation Data (2016).